The Future of Serials in a Linked Data World

Laurie Kaplan
Serials Solutions, a ProQuest business, lkaplan@proquest.com

Laurie Kaplan, "The Future of Serials in a Linked Data World" (2012). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284315127

The Future of Serials in a Linked Data World

Laurie Kaplan, Director of Editorial Operations, Serials Solutions

Author Note

Laurie Kaplan is the Director of Content Operations for Serials Solutions, a ProQuest business. Laurie is also currently an MLIS student at Rutgers, the State University of New Jersey. A version of this paper is being submitted as a term project for her Digital Libraries course for the 2012 fall semester.

This paper is based on a presentation prepared by Valerie Bross and Laurie Kaplan, and presented at the 2012 Charleston Conference by Valerie Bross and Yvette Diven (in Laurie's absence). Valerie's review of this paper contributed to the final version. Valerie Bross is the Head of the Continuing Resources Cataloging Section, UCLA Library Cataloging and Metadata Center and Yvette Diven is Publisher and Senior Product Manager for Serials Solutions, a ProQuest business.

Abstract

Serials, from a cataloging, search, and retrieval point of view, are currently described and accessed via metadata records. Each record is tied to the title of the journal, newspaper, or magazine. The record might cover a range of years for that publication under its current title, or it might cover the current iteration and previous titles. But in our libraries, to find a serial we look for the appropriate record, usually a MARC record, in OPACs and search systems. The cataloging rules are changing, and RDA will soon replace AACR2 as the content standard for creating MARC records and other library metadata for books and serials. The Library of Congress has announced that as the cataloging rules are changing, so too will the bibliographic framework change. The current framework, FRBR (a linear, hierarchical conceptual model) and the MARC standard (the flat format used for catalog records in the US and many other countries around the world), form the basis of many catalog records. All signs are pointing toward a new framework built on RDF and linked data. How will current MARC records adapt to use in a linked data world? Should future structures and displays use the traditional hierarchical approach, or should they take as a model the web-like structure taking shape for the Semantic Web? And how can libraries and librarians take part in this next phase of information access and retrieval?

A hot topic in today's literature and at library and information science conferences is linked data. Everyone wants to be part of the linked data world, and it is referred to as a new concept. It may be a new concept to electronically link disparate content, but many of the principles of linked data have been applied by librarians for as long as there have been libraries, including the classification of data and making resources accessible to library patrons. In ancient Egypt around 300 B.C., the Library of Alexandria used a classification system for their papyrus scrolls and arranged them in bins by subjects. In the United States in the 1770s, Thomas Jefferson classified his personal library by subject and chronology, using broad subjects such as Science, Memory (History), Reason (Philosophy), and Imagination (Fine Arts). About 100 years later, three classification systems were developed, each with varying degrees of detail and granularity. The Dewey Decimal System (1876), the Cutter Classification System (1882), and the Library of Congress Classification System (1897) all created classifications and enabled patrons to find linked data on the shelves in the library. Without that cataloging and classification, how would a patron find the history section with books and periodicals specifically about the Renaissance, and even more specifically about Italian art, Michelangelo, and Leonardo da Vinci?

Librarians in more modern times have been using the Anglo-American Cataloging Rules, 2nd Edition (AACR2) and MARC (Machine Readable

http://dx.doi.org/10.5703/1288284315127

Cataloging) to create catalog records for the materials in their libraries. Among the most challenging records are the serials records, tracking their title changes as they merge, incorporate, and split apart again over their lifespans. Serials include several types of publications as well, such as scholarly journals, newspapers, government documents, consumer and trade magazines, annuals, reports, yearbooks, directories, proceedings, and monographic series. Additionally there are self-published family newsletters, 'zines, and online publications. In order to keep up with the changing face of library catalogs and with electronic library systems, cataloging rules are changing as well with Resource Description and Access (RDA) replacing AACR2. The Joint Steering Committee for the Development of RDA (2012) notes, "RDA provides a set of guidelines and instructions on formulating data to support resource discovery…covering all types of content and media." One question to be explored is how librarians can support linked data on the Semantic Web and how that will change the way librarians catalog and classify resources. The Semantic Web is defined by the World Wide Web Consortium (W3C) as "a Web of data—of dates and titles and part numbers and chemical properties and any other data one might conceive of. RDF, which stands for Resource Description Framework, provides the foundation for publishing and linking … data." It is "a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time…."

The intention of this paper is to review the literature on the concept of linked data, relate linked data to serials publishing, and discover what librarians are currently planning and what they might do in the near future to facilitate the creation of linked data. The methods employed are a review of the current research, websites, and linked data models to analyze which of these are applicable to serials publications and serials research. This paper will also discuss several sites that can be used by serials librarians to create or enhance metadata in their catalog records. The objective of this research is to understand the implications of linked data for serials librarians and to understand how serials librarians can take advantage of enhanced navigation between the traditional online library resources and the rest of the web through the use of metadata and linked data to help their patrons improve research results in a linked data world.

Current Landscape of Serials—Access to Research

Serials are a major part of the library collection in any format imaginable, including print publications, online publications, open access collections, and institutional repositories, with or without their accompanying datasets. One critical role of the librarian is to make these resources accessible for their patrons, researchers, students, and the general public, through catalog records and a search system. Technological advances in library research are dependent upon the underlying metadata and authorities to ensure that search terms employed by patrons result in finding the appropriate resources. There are many standards for expressing that metadata, some of which are general standards for bibliographic records (MARC and Dublin Core), and others that are designed for specific types of records (TEI and EAD), and all of which can now be included in more detail thanks to RDA, the replacement for AACR2 bibliographic description.

There are also new standards and identifiers for adoption, including:

- International Standard Name Identifier (ISNI): an ISO Standard (ISO 27729:2012) whose scope is the identification of public identities across multiple fields of creative activities, disambiguating natural, legal and fictional parties that might otherwise be confused;

- International Standard Text Code (ISTC): a numbering system for the unique identification of text-based works; the term "work" can refer to any content appearing in conventional printed books, audio-books, static e-books or enhanced digital books, as well as content which might appear in a newspaper or journal; and

- Open Researcher and Contributor Identifier (ORCID): somewhat similar to ISNI, intended to disambiguate author names, starting with scholarly journal authors first, and linking to scholarly object identifiers.

RDA, RDF, and Bibliographic Framework

Most current US catalog records use AACR2 as their bibliographic description or content formalization, and are most often encoded using MARC21 format. The move to RDA from AACR2 has raised issues about the long-term viability of the MARC format. One reason is that MARC does not represent relationships and hierarchies between pieces of bibliographic data, which is a feature of RDA, following a more web-like model and identifying and relating the resources in library collections. As a result, the Library of Congress announced the Bibliographic Framework Initiative in October, 2011 to investigate alternatives: "The new bibliographic framework project will be focused on the Web environment, Linked Data principles and mechanisms, and the Resource Description Framework (RDF) as a basic data model." According to a presentation given by Sally McCallum (2012), some of the requirements for the Bibliographic Framework Initiative are enhanced linking for semantic technology through Uniform Resource Identifiers (URIs), MARC compatibility with the continued maintenance of MARC21 and the ability to reuse MARC data, and new views of different types of metadata (e.g., descriptive, authority, holdings, classification, subject, rights). The initiative will use the Web as a model for connecting information, and will investigate the use of the RDF data model and various syntaxes in a collaborative way. The linked data orientation will lead to easier integration of catalog data with data on the web and in social media, increase flexibility for descriptive data, and facilitate reuse of data for searching and applications. McCallum (2012) noted that while balancing factors they would "leverage machine technology for the mechanical while keeping the *librarian* expertise in control." Kevin Ford, project manager for the Library of Congress Linked Open Data service, noted that "RDF provides a means to represent the data and the Linked Data methods and practices provide a means to communicate the data, the two core and historical functions of MARC" (Ford, 2012, p. 46). He further noted that "Linked Data is about publishing structured data over the same protocol used by the World Wide Web and linking that data to other data to enhance discoverability of more information" (Ford, 2012, p. 47).

The process to move from MARC to a new linked data model will be gradual to enable librarians to manage their legacy data and incorporate it into the linked data world. While the changes apply to resources generally, they are particularly challenging in the case of serials. Classic serials issues of tracking title changes over time, finding the appropriate copy, retrieving all parts of the serial, including articles, bibliographies, graphs, and images, are multiplied when moving from one system to another. Of utmost importance to serials librarians is ensuring that serials are properly coded so that systems used for search and retrieval can successfully resolve all the links and find the results for students and researchers.

Current Research on Linked Data

There are many articles about linked data available on the web. A search for "linked data" (with the quotes) in two library systems using Serials Solutions® Summon™ service finds 7,947 results in one and 8,968 in the other; the same search in Google finds 16,500,000 results and in Google Scholar finds 26,500 results. Several websites and articles credit Tim Berners-Lee, the current Director of the World Wide Web Consortium (W3C), with coining the term "linked data" in his 2006 Design Notes. In that same document, he defines the four principles (or "expectations of behavior"), which were summarized by Bizer, Heath and Berners-Lee (2009) as

- Use URIs as names for things,
- Use HTTP URIs so that people can look up those names,
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL), and

- Include links to other URIs, so that they can discover more things.

These principles set the stage for publishing data on the web and for connecting data via the web within a framework or structure and standards. The basic grammatical structure of the RDF model is stated "in the form of *subject, predicate, object* triples. The subject and object of a triple are both URIs that each identify a resource….The predicate specifies how the subject and object are related, and is also represented by a URI" (Bizer, Heath, & Berners-Lee, 2009). Bizer, et al. (2009) also note that "it is possible to think of RDF triples that link items in different data sets as analogous to the hypertext links that tie together the Web of documents."

Needleman (2007) noted that "RDF allows for both human-readable and machine-parseable vocabularies and is designed to support the reuse of metadata semantics and vocabularies among disparate information communities" (p. 58). Much of linked data involves markup languages and vocabulary standards. The most common mark-up languages are HTML and XML, and RDF works with both. "RDFa is an extension to HTML5 that helps you markup things like People, Places, Events, Recipes and Reviews. Search Engines and Web Services use this markup to generate better search listings and give you better visibility on the Web, so that people can find your website more easily….In fact, if your markup language is based on XML, then you can already use RDFa in your documents today" (Linked data in HTML). According to a blog post by Eric Hellman (2009) "there is an easy and rough transformation to go from marc into an RDF model: the triples are (record ID, marc field/subfield, field value). A single MARC record decomposes into many triples." As we are beginning to see, standards that are already in use are relevant for linked data, and can provide structure for data and objects on the web, including the various types of serials publications.

Examples of Linked Data in Library Settings

The Library of Congress Linked Data Service was created in early 2009, with 17 datasets now available; the list can be found at http://id.loc.gov/descriptions/. "The Library of Congress Linked Data Service enables both humans and machines to programmatically access authority data at the Library of Congress….The Library of Congress has prepared this vocabulary terminology system and is making it available as a public domain data set." Also available since 2009 is the Swedish National Library's Union Catalog published as linked data; similar efforts from the German and French national libraries, and the British Library followed over the last few years.

OCLC is involved with several linked data initiatives including xISSN, Dewey Web Services, VIAF (Virtual International Authority File), and the Schema.org initiative with WorldCat.org, Google, Bing, Yahoo, and Yandex, in a "cooperative agreement between these major search engines to share a core vocabulary for markup" (Fons, Penka, & Wallis, 2012, p. 29). The xISSN web service project was one of the earliest to use linked data through standard numbers. It pulls together associated serials based on the ISSN, is machine actionable, and offers a visualization of the serial map that allows for human interpretation. Dewey Decimal Classification (DDC) top three levels became available as linked data in September 2009, and the summaries are now also available as linked data from dewey.info (June 2012). Adding RDF vocabulary and URIs to the summaries extends their web document version, and enables anyone using Dewey numbers to add URIs and link to the summaries, available in nine languages. Updates are also automatically available through the links. Another project also launched in September 2009, and now hosted by OCLC is VIAF.org, is "a joint project of several national libraries plus selected regional and trans-national library agencies. The project's goal is to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web" (VIAF, http://viaf.org/). VIAF creates a "super" authority record by linking together all authority data for a given entity.

The newest OCLC project is the OCLC Linked Data Initiative with Schema.org, which was also released in June of 2012 for over 250 million

records on WorldCat.org. This project is adding "a set of vocabulary extensions to WorldCat data. Schema.org and library-specific extensions will provide a valuable two-way bridge between the library community and the consumer web" (Fons, et al., 2012, p. 30).

A final wide-ranging group to mention is the Linked Open Data in Libraries, Archives, and Museums (LODLAM) Summit which was founded to help educate the global community of these organizations about linked data and its potential for making connections among disparate datasets throughout the world. Their mission, initiated at the first meeting in June of 2011, is to foster discussion on issues of metadata, vocabularies, copyright, licensing, and more and to encourage collaborative projects of linked open data to demonstrate the successes and the issues. LODLAM continues to encourage discussion and presentations at their annual meetings, regarding linked data for archives, museums, and all library resources, including serials and monographs.

The Future—Issues for Libraries, Publishers, and Vendors

The move to linked data is not without issues; legal issues of copyright and licensing and differences in underlying metadata formats and languages need to be resolved before linking can proceed. And as far as they have come, there is still more to be done, including improved connectivity, developing standards, ensuring interfaces that enhance navigation, and improving integration. The OCLC projects also have goals for the future, including improving and adding vocabularies to extend the basics of Schema.org, improving access to data, and extending the links that are currently mapped.

Karen Coyle, a frequent author on the topic of library data, notes that the design of useful data has changed over time from an alphabetic card catalog system, through keyword searching, to facets for more narrow or focused searching, to linked data. For the first three, there needed to be additions to a library record or a method to manage the library catalog, usually in a database. "Designing data for linking goes beyond additions to the catalog record: it requires that we adopt a significantly different metadata methodology. This methodology is based on technologies that permit sharing data over the web and making connections between disparate data stores based on the data elements that they have in common" (Coyle, 2011, p. 156). She notes that linked data uses structured statements for data and metadata, controlled vocabularies whenever possible, and identifiers to name every item. The move to linked data from standard catalog records requires an analysis of the existing data, ensuring that as much of the text as possible is converted to data (so it can be analyzed), which likely requires breaking apart existing fields to restructure as RDF or similar structured data. One example is in catalog entries for standard numbers such as ISSN. If the term Online or Print is added to that field, the identifier becomes textual, not data. In order for a machine to understand the data, it has to ignore the text comment. Many libraries are undertaking this analysis now in preparation for a move to a linked data catalog. "It is definitely not a matter of serving only the machine or only the human reader, but of creating data that can serve both" (Coyle, 2010, p. 15).

Laura Krier (2012) "proposes a new way of cataloging serials using linked data and Resource Description Framework (RDF), as well as how the concepts of Functional Requirements for Bibliographic Records (FRBR) can be expanded to apply to journal content at both the journal level and the article level, all with an eye toward ease of access and understanding for users" (p. 177). The move to linked data would require "an item to be cataloged as a resource is assigned a URI that is available on the open Web. A cataloger would then use element sets such as the Dublin Core Metadata Initiative terms, the International Standard Bibliographic Description (ISBD) terms, or FRBR concepts in RDF to describe that resource by making statements about it" (Krier, 2012, p. 180). In the absence of separate records, a cataloger would add statements that link the item, wherever it is, to the library to indicate that it is included in the library's collection. Specific local notes can be published and different users can pull elements to suit their needs. Krier (2012) believes that the complex nature of serials and

their bibliographic relationships would work well with the linked data model. Linked data can minimize the complications of serials title changes and multiple formats by focusing on the links to resources rather than the description of those resources, and the focus would also shift to the electronic resources as more of them become available, including digitized back issues. According to Krier (2012) "the shift to a linked data model would not only help users better understand the bibliographic universe; it would save immense amounts of time for catalogers, too…catalogers can work collaboratively to maintain bibliographic metadata, and take advantage of metadata released by publishers and vendors" (p. 185).

Providers of data for libraries are also beginning to add RDF and URIs to the underlying markup, to enable linking of that data. The Directory of Open Access Journals (DOAJ) has "exposed the data behind their system. Normally one would need to click several times to get to the data one is looking for, such as a specific title" (Miller, 2012, p. 20). Using the exposed data from DOAJ and connecting it with visualization and navigation tools can enable a library to combine that data with other data in building a collection development strategy.

Additional issues to consider are whether linked data can improve the identification and mapping of serials title changes to ensure that the appropriate copy is available regardless of the citation or standard number identifier used; aid in the curation of data sets, multi-media files, and other research data that inform the article, journal or project to enable access to these elements; and improve link resolution for citations to materials held in the collection and available on the web.

The Ongoing Role of Librarians

There is no doubt that librarians will continue to talk about, blog about, and conduct research about linked data and will share their thoughts, trials, and tribulations at conferences, on professional organization and user group lists, and through social media. The persistence of librarians to ensure an organized move from MARC to linked data through the Bibliographic Framework Initiative, their advocacy for change from data and systems providers, and their experimentation with linked data for their own library collections will lead the way for serials in the linked data world (Byrne & Goddard, 2010). As noted in the beginning of this paper, librarians have been working with linked data for as long as there have been libraries. To ensure the continued relevance of the data in library collections, librarians will begin to enhance those collections using Semantic Web technology and adding metadata and URIs to thesaurus, mapping, and taxonomy services. There will be a shift away from focusing on records to focusing on data and ensuring that the relevant pieces of data are represented by metadata that can be found on the web. Library catalogs hold a wealth of data about published and unpublished materials. Adding that data to the linked data cloud or universe will increase the search success rate for a researcher, and combined with other linked data on the web may lead the researcher to some new observations and conclusions. Transformation of serials data is needed, and the existing MARC tags and metadata elements are a great beginning.

References

Berners-Lee, T. (2006). Linked data. *W3C Design Issues.* Retrieved from http://www.w3.org/DesignIssues/LinkedData

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—The story so far. *International Journal on Semantic Web and Information Systems, 5*(3), 1–22.

Byrne, G., & Goddard, L. (2010). The strongest link: Libraries and linked data. *D-Lib® Magazine, 16*(11/12). Retrieved from http://www.dlib.org/dlib/november10/byrne/11byrne.html

Coyle, K. (2010). Chapter 2: Changing the nature of library data. *Library Technology Reports, 46*(1), 14–29. Retrieved from http://alatechsource.metapress.com/content/P7W420J45503JK0W

Coyle, K. (2011). Designing data for use: From alphabetic order to linked data. *Serials, 24*(2), 154–159. http://dx.doi.org/10.1629/24154

Coyle, K. (2012). Linked data tools: Connecting on the web. *Library Technology Reports, 48*(4).

Dewey Services: Dewey Web Services. (n.d.). Retrieved from http://www.oclc.org/dewey/webservices/default.htm

Fons, T., Penka, J., & Wallis, R. (2012). OCLC's linked data initiative: Using schema.org to make library data relevant on the web. *Information Standards Quarterly (ISQ), Spring-Summer,* 29–33. http://dx.doi.org/10.3789/isqv24n2-3.2012

Ford, K. (2012). LC's bibliographic framework initiative and the attractiveness of linked data. *Information Standards Quarterly (ISQ), Spring-Summer,* 46–50. http://dx.doi.org/10.3789/isqv24n2-3.2012

Harper, C. (Ed.). (2012). Linked data in libraries, archives, and museums [Special issue]. *Information Standards Quarterly (ISQ), Spring-Summer*. http://dx.doi.org/10.3789/isqv24n2-3.2012

Hellman, E. (2009, July 8). Re: Yee on RDF and bibliographic data [Weblog post]. Retrieved from http://kcoyle.blogspot.com/2009/07/yee-on-rdf-and-bibliographic-data.html

Joint Steering Committee for the Development of RDA. (n.d.). Retrieved from http://www.rda-jsc.org/rda.html

Krier, L. (2012). Serials, FRBR, and library linked data: A way forward. *Journal of Library Metadata, 12*(2–3), 177–187. http://dx.doi.org/10.1080/19386389.2012.699834

Linked Data in HTML. (n.d.). Retrieved from http://rdfa.info

Marcum, D. (2011). Bibliographic framework transition initiative. *Library of Congress.* Retrieved from http://www.loc.gov/marc/transition/news/framework-103111.html

McCallum, S. (2012). Bibliographic framework initiative approach for MARC data as linked data [PowerPoint slides]. Retrieved from http://www.loc.gov/marc/transition/

McGrath, K. (January, 2011), Will RDA kill MARC? [PowerPoint slides]. Retrieved from http://pages.uoregon.edu/kelleym/KM_MWpresentation.pdf

Miller, E., & Westfall, M. (2011). Linked data and libraries. *The Serials Librarian, 60*(1–4), 17–22. http://dx.doi.org/10.1080/0361526X.2011.556427

Needleman, M. (2001). RDF: The resource description framework. *Serials Review, 27*(1), 58–61.

Tillett, B. (January, 2010). RDA changes from AACR2 for texts [PowerPoint slides]. Retrieved from http://www.rda-jsc.org/docs/10_1_12_RDAchangesfromAACR2fortexts.ppt

Voss, J. (2012). LODLAM state of affairs. *Information Standards Quarterly (ISQ), Spring-Summer*. http://dx.doi.org/10.3789/isqv24n2-3.2012

Wilson, B., & Fenner, M. (May, 2012). Open Researcher and Contributor ID (ORCID): Solving the name ambiguity problem. *EDUCAUSEreview online.* Retrieved from http://www.educause.edu/ero/article/open-researcher-contributor-id-orcid-solving-name-ambiguity-problem

World Wide Web Consortium. (2012). Retrieved from http://www.w3.org/rdf/ and http://www.w3.org/standards/semanticweb/